MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 10: 105–129 (2026)
DOI: 10.3897/mbmg.10.164232

Research Article

Filter pore size influences taxonomic composition of retained eDNA from seawater samples—evidence from shotgun sequencing

Adrián Gómez-Repollés[1,2], Eva Egelyng Sigsgaard[1], Mads Reinholdt Jensen[1,3], Philip Francis Thomsen[1]

1 *Department of Biology, Aarhus University, Ny Munkegade 116, Building 1540, DK-8000 Aarhus C, Denmark*
2 *Department of Molecular Biology and Genetics, Aarhus University, Universitetsbyen 81, Building 1872, DK-8000 Aarhus C, Denmark*
3 *Norwegian College of Fishery Science, UiT – The Arctic University of Norway, NCFS-building (NFH), NO-9037 Tromsø, Norway*
Corresponding author: Philip Francis Thomsen (pfthomsen@bio.au.dk)

Abstract

Environmental DNA (eDNA) is increasingly used in aquatic research and applied biodiversity monitoring, mainly through a metabarcoding approach. This allows detailed analyses of specific taxa of interest but involves amplification by PCR, which can lead to significant taxonomic bias. On the other hand, broad tree-of-life approaches such as shotgun sequencing are easily dominated by prokaryotic DNA, failing to detect less abundant taxa such as metazoans. However, the proportion of eukaryotic DNA could potentially be increased through water filtration with larger pore sizes. In this study, we investigate how filter pore size influences the taxonomic composition of retained aquatic eDNA. We conducted high-throughput shotgun sequencing and 18S rDNA metabarcoding on 15 seawater eDNA samples collected at the same locality but filtered with different pore sizes spanning from 0.2 to 8.0 µm. We could assign only 0.78% of the total shotgun reads to the superkingdom level, but smaller pore sizes (0.2 and 1.2 µm) yielded significantly higher proportions of bacterial reads compared to eukaryotic reads (63 vs. 28%), while larger pore sizes (5.0 and 8.0 µm) yielded significantly higher proportions of eukaryotic reads compared to bacterial reads (49 vs. 31%). Of the 19 metazoan phyla detected with shotgun sequencing, all but one were more abundant in larger pore sizes. Most of the eukaryotic phyla detected (39 of 54) were shared between shotgun and metabarcoding analyses, indicating similar performance of the methods regarding presence/absence estimation at high taxonomic levels. Finally, we compared the reliability of genus-level detections between the two methods for selected groups of metazoans to evaluate the potential of shotgun sequencing for biomonitoring. Our study suggests that the prokaryotic dominance in shotgun sequencing data from aquatic eDNA can be reduced by increasing filter pore size. As reference databases expand, the taxonomic coverage and resolution of shotgun sequencing should improve, likely enhancing the potential of shotgun sequencing for future eDNA research.

Key words: Biodiversity, eukaryote, marine, metabarcoding, metazoan, Sterivex

OPEN ACCESS

Academic editor: Dominik Buchner
Received: 5 July 2025
Accepted: 12 January 2026
Published: 18 February 2026

Citation: Gómez-Repollés A, Sigsgaard EE, Jensen MR, Thomsen PF (2026) Filter pore size influences taxonomic composition of retained eDNA from seawater samples—evidence from shotgun sequencing. Metabarcoding and Metagenomics 10: e164232. https://doi.org/10.3897/mbmg.10.164232

Copyright: © Adrián Gómez-Repollés et al. This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Introduction

Over the past two decades, environmental DNA (eDNA) analyses have increasingly been applied to aquatic systems, both to detect single species and to identify biological community compositions (reviewed in Takahashi et al. 2023). The development of high-throughput sequencing (HTS) technologies in conjunction with plummeting sequencing costs has been key to the emergence of eDNA analysis as a supplementary tool in both fundamental and applied research, as well as for biomonitoring of both micro- and macro-organisms. The applications of eDNA are diverse, ranging from analyzing community compositions (Agersnap et al. 2022) to estimating population genetic variation (reviewed in Sigsgaard et al. 2020). Across these applications, eDNA metabarcoding is the most widely applied approach and has proven highly effective for studying specific taxa of interest, but an important drawback is that it requires PCR amplification. This inherently limits the taxa that can be detected and makes quantitative estimates difficult, as target taxa are often unequally amplified (Pan et al. 2014; Shirazi et al. 2021; Silverman et al. 2021). In recent years, shotgun sequencing of eDNA samples has been increasingly explored to avoid the issues associated with PCR bias. Metagenomic approaches using shotgun sequencing have long been used to characterize microbial communities (Quince et al. 2017) but have only recently been applied to macro-organisms. Here, it has shown promising results in ancient eDNA studies (Kjær et al. 2022) and has also been applied to modern eDNA samples, where it has been used to target single species through target capture enrichment protocols (Pinfield et al. 2019; Jensen et al. 2021). It has also been used to target eukaryotes more broadly. For example, Cowart et al. (2017) detected organisms across the tree of life in the West Antarctic Peninsula using shotgun sequencing of eDNA samples. Similarly, Stat et al. (2017) applied shotgun sequencing to characterize marine biodiversity in a tropical marine environment. However, the great majority of sequencing reads in both studies came from bacteria, and Stat et al. (2017) highlighted the method's inefficiency for eukaryotes when compared with metabarcoding.

Currently, there are no consensus guidelines to follow for aquatic eDNA sampling, even though multiple studies define recommendations or practical guides (Andruszkiewicz et al. 2017; Goldberg et al. 2016), and a wide range of different filter types and filter pore sizes are used to collect eDNA samples from water (Bowers et al. 2021). The enclosed Sterivex filter is one of the most widely used filter types (Miya et al. 2015; Sigsgaard et al. 2017; Antich et al. 2021) due to the reduced risk of contamination and ease of use. Nonetheless, this filter is relatively expensive and is available only with pore sizes below 1.0 µm, which can lead to clogging (e.g., Turner et al. 2014). For these reasons, other alternatives, such as unenclosed filters of varying pore sizes, are also frequently used (Takahashi et al. 2023), which, on the other hand, come with the caveat of increased risk of contamination.

Filter pore sizes are expected to influence results, since eDNA may be present in many different states, including but not limited to complete organisms, sloughed tissue, feces, free DNA, or gametes (Thomsen and Willerslev 2014; Turner et al. 2014; Tillotson et al. 2018; Mauvisseau et al. 2022; Power et al. 2023), and since cell sizes differ between taxonomic groups (Fowler et al.

2017). For metazoans, the most abundant eDNA particles are from 1 to 10 µm in size, corresponding to or larger than the size of whole organelles and cells (Power et al. 2023). However, studies of aquatic eukaryotic biodiversity using eDNA have applied filters with pore sizes ranging from 0.2 µm up to 180 µm (e.g., Turner et al. 2014), with filter pore sizes below 1 µm being the most common (reviewed in Takahashi et al. 2023). With current efforts to apply PCR-free or near PCR-free analyses to eDNA monitoring, using only a few PCR cycles, investigating the effect of filter pore size across metabarcoding and shotgun is important. Filter pore sizes are expected to impact the ratio of microbial to macrobial eDNA captured (Jo et al. 2022), but this remains largely untested. Pascoal et al. (2023) investigated the effect of both water volume and filter pore sizes on prokaryotes and single-celled eukaryotes using shotgun sequencing and metabarcoding. Here, a strong effect of pore size was found for both richness and composition of species, but the microbial-to-macrobial DNA ratio was not investigated. Efficiency of the metagenomic method might well be improved for eukaryotes by increasing filter pore size.

In this study, we perform parallel high-throughput shotgun sequencing and metabarcoding of seawater eDNA samples filtered with pore sizes spanning 0.2 to 8.0 µm to test the hypothesis that larger pore sizes will result in a higher proportion of eukaryote eDNA. Additionally, we investigate how well shotgun sequencing covers phylum-level eukaryotic diversity compared to the more commonly used 18S rDNA metabarcoding. While both the rates of taxonomic assignment (proportion of sequences assigned to at least the superkingdom level) and the taxonomic resolution of assigned sequences were expected to be lower for shotgun sequencing than for metabarcoding, we expected that shotgun sequencing might yield a broader taxonomic coverage at high taxonomic levels.

Materials and methods

Study site and sampling

Water sampling for eDNA analysis was conducted at Skovshoved Harbour (Copenhagen, Denmark, 55.76386°N, 12.60299°E) in October 2017. We filtered 15 samples in total, each consisting of 1 L of surface seawater, using a large 60 mL syringe (Soft-Ject, HSW, Tuttlingen, Germany), which was reused but rinsed between each sampling. Triplicate samples were collected with each of five different combinations of filter pore size and filter type: four different pore sizes (0.2, 1.2, 5.0, and 8.0 µm) of sterile non-enclosed (47 mm diameter; filtration area: ca. 17 cm^2) filters (Sterlitech, USA) and one pore size (0.2 µm) of the enclosed (filtration area: ca. 10 cm^2) filter Sterivex-GP (Merck Life Science, Denmark). All filters were made of polyethersulfone (PES). The applied pore sizes were chosen based on the expected range of cell sizes of Eukaryota (10–100 µm) compared to Bacteria (0.1–5.0 µm) (Fowler et al. 2017), such that the two largest pore sizes were expected to capture a higher proportion of eukaryotes. To avoid biases due to temporal changes in eDNA composition at the sampling site, one replicate was taken of each filter type before moving on to the next replicate, and the time gap between samples was minimized as much as possible. Within each of the three replicate sampling sessions (named A,

B, and C), the sampling was ordered with the non-enclosed (hereafter "OP" for open) (0.2, 1.2, 5.0, and 8.0 µm) filters first, going from the smallest pore sizes to the largest, followed by the enclosed filter (Sterivex-GP 0.2 µm, hereafter "EN0.2" for enclosed). The sampling process was continuous and lasted ca. 3 hours, with approximately one minute elapsed between samples. Sampling and filtering were done on-site using facemasks and nitrile gloves to reduce contamination. Filters were kept cold with cooler packs during fieldwork and transport. Later on the same day, after returning from the field, the filters were frozen and kept at –20 °C until DNA extraction.

DNA extraction

Extractions of eDNA from filters were performed in the clean laboratory facility at the Department of Biology, Aarhus University. This is a laboratory dedicated to eDNA and museum specimens, where fresh tissue samples are not allowed, and a range of anti-contamination protocols are followed. The DNA was extracted as in Jensen et al. (2022), using the DNeasy® Blood & Tissue kit (Qiagen) with a modified version of the manufacturer's protocol. Sterivex filters were cleaned with a 20% dilution of household bleach followed by 70% ethanol. A lysis mix consisting of 80 µL of proteinase K and 720 µL of ATL buffer was added to each filter, and the filters were then incubated for 3 hours at 56 °C while being agitated on a rotor. The lysis mix was transferred from the filters to a 5 mL Eppendorf tube and mixed with 800 µL of absolute ethanol and 800 µL of AL buffer. The tubes were then vortexed thoroughly before the mix was spun through a spin column, 700 µL at a time. The spin columns were then washed with 500 µL of AW1 buffer followed by 500 µL of AW2 buffer. Finally, the DNA was eluted over two rounds, each with 60 µL AE buffer and 10 minutes of incubation at 37 °C. An extraction blank was included for each extraction setup and subjected to the same protocol as the samples, except that smaller volumes were used (180 µL ATL buffer, 20 µL proteinase K, 200 µL AL buffer, and 200 µL ethanol). All DNA extracts were stored at –20 °C.

Library building and shotgun sequencing

The extracted DNA was sent to Novogene (Cambridge, UK) for library preparation and shotgun sequencing. The extraction blanks were not sequenced due to DNA concentrations being too low for library building. Libraries were built using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, England), following the manufacturer's recommendations, which include 3–15 PCR cycles. The genomic DNA was randomly fragmented to a size of 350 bp with a Bioruptor, and the DNA fragments were size selected with sample purification beads. The selected fragments were then end-polished, A-tailed, and ligated with the full-length adapter. After these treatments, the fragments were size selected with beads again. Lastly, the library was analyzed for size distribution with an Agilent 2100 Bioanalyzer and quantified using real-time PCR. Quality control of the libraries, based on DNA degradation or contamination, concentration, and fragment size distribution, was verified using agarose gel electrophoresis, Qubit, and a TapeStation (Agilent), respectively. The libraries were sequenced on an Illumina NovaSeq 6000 platform using 150 PE and

requesting 10 gigabase pairs (Gbps) of output per library, where at least 80% had a quality score (Q score) equal to or above 30. After sequencing, reads containing adapters, reads containing > 10% undetermined bases (Ns), and reads where > 50% of the bases had a Q score <= 5 were removed by Novogene.

PCR amplification and library building for metabarcoding

We performed PCR targeting eukaryotes using the primers 18S_allshorts forward (5'-TTTGTCTGSTTAATTSCG-3') and 18S_allshorts reverse (5'-CACAGACCTGTTATTGC-3'), which amplify approximately 110 bp of the 18S rRNA gene (Taberlet et al. 2018). Compared to the original version of the reverse primer (Guardiola et al. 2015), the version in Taberlet et al. (2018) lacks a "T" at the 5'-end to better equilibrate the melting temperatures of the forward and reverse primers. Using 17 different primer twin-tags (consisting of 2–3 N's (De Barba et al. 2014) and a unique sequence of six nucleotides designed with the software OligoTag (Coissac 2012), we individually tagged the 15 eDNA samples and the two extraction blanks. A final tag was used for all PCR blanks. Four PCR replicates were run for each sample and blank. The PCR setups were run in 25 µL volumes, using 10 µL HotStarTaq Master Mix (Qiagen, Cat. no. 203445), 10 µL ddH_2O, 1 µL primer mix (5 µM forward and 5 µM reverse), 1 µL BSA (Bionordica, Cat. no. B9000S), and 3 µL DNA template. The thermocycler conditions were set to an initial 10 minutes of denaturing at 95 °C, followed by 45 cycles of 94 °C for 30 s, 45 °C for 30 s, and 72 °C for 30 s. Product lengths were verified on 2% agarose gels stained with GelRed™. The PCR products were combined into four pools, each containing one PCR replicate (6 µL volume) per sample and blank. The pools were then purified using the MinElute PCR Purification Kit (Qiagen). We followed the manufacturer's protocol with the exception that samples were incubated with the elution buffer (20 µL) at 37 °C for 10 minutes. A library was prepared from each pool using the TruSeq DNA PCR-Free LT Sample Prep Kit (Illumina, San Diego, California), with an input of ca. 750 ng of purified PCR product from each pool, as measured with a Qubit HS DNA Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). A PCR-free library kit was used, as this minimizes the risk of tag jumping (Schnell et al. 2015). Libraries were then sequenced on the Illumina NovaSeq 6000 platform by Novogene using 150 PE sequencing and requesting 10 Gbp of output per library.

Data analysis for shotgun sequencing

Raw read quality control was performed using FastQC (v0.11.8) before and after merging the forward and reverse reads. We used BBMap (v38.72) (Bushnell 2014) to merge the reads, defining a minimum overlap of 35 bp, a maximum of 5 mismatches, and a minimum insert size of 40 bp. Fastq files were transformed to fasta format using seqtk (v1.3-r106) and were then split into smaller files, each consisting of 5,000 reads, using the Bash function split (v8.22) to allow parallel BLAST searches. We used BLASTn (v2.5.0+) (Altschul et al. 1990; Benson 2004) to compare the reads against a local copy of the GenBank nt database (version: 27 October 2019; downloaded on 30 March 2020), allowing for a maximum of 1,000 hits per read. We matched the

BLASTn results and the "categories.dmp" file, obtained from the NCBI "taxdump" to identify reads at the superkingdom taxonomic level (A: Archaea; B: Bacteria; E: Eukaryota; V: Viruses and viroids; U: Unclassified; O: Other). We merged reads belonging to "Other" with the category "Unknown" and reads matched to multiple categories were labeled as "Ambiguous". Following this, we explored the effect of the BLASTn similarity threshold on taxonomic identifications by using increasing percentages of similarity (70% to 95% in increments of 5%, 99%, and 100%) to identify a reliable similarity cutoff (Suppl. material 1: fig. S1).

A table containing the full taxonomic classification (superkingdom, kingdom, phylum, class, order, family, genus, and species) from NCBI for each of the taxids present in the database hits was prepared using the R package taxizedb (v0.3.0) (Chamberlain and Arendsee 2021). Using an in-house script, reads whose best hits had an e-value of > 0.0001 (1/10,000), a sequence similarity of < 90%, a query coverage of < 90%, or an alignment length of < 100 bp were discarded, and the remaining reads were taxonomically identified using the previously prepared taxonomic table. Reads with hits to multiple superkingdoms were assigned to the predominant one when at least 90% of their hits belonged to that superkingdom. Once assigned, hits to other superkingdoms were discarded. If this 90% dominance threshold was not met, reads were considered ambiguous and excluded from further analysis. The remaining reads were then used to identify the lowest common ancestor (LCA) without additional filtering. To account for differential library sequencing depth, the resulting data were rarefied to the median value of the sample read counts (de Cárcer et al. 2011) using the R package ROBITools (v0.1) (https://git.metabarcoding.org/obitools/ROBITools). After rarefying, taxa with zero reads were discarded. We then used the function metaMDS from the R package vegan (v2.6.4) (Oksanen et al. 2020) to perform non-metric multidimensional scaling (nMDS) using the Bray–Curtis and Jaccard dissimilarity indices on read counts and presence/absence of phyla, respectively. Results of nMDS were visualized using the package tidyverse (v2.0.0) (Wickham et al. 2019). Finally, zero-handled read counts (value + 1) were transformed using the centered log-ratio (CLR) function from the package compositions (v2.0-6) and visualized using the R package ComplexHeatmap (v2.14.0) (Gu et al. 2016) to compare read enrichment or depletion, relative to the individual geometric mean per taxon, across pore sizes.

We used contingency tables of rarefied read counts to calculate, for each filter pore size, the odds ratio (OR) for obtaining a read from Bacteria and Eukaryota, respectively, relative to a read from another superkingdom. Similarly, we calculated the OR for obtaining reads from each superkingdom in the three different sample replicates relative to each other (combinations: A vs. B + C; B vs. A + C; C vs. A + B; C vs. A; and C vs. B). We performed Wilcoxon tests using the stats package in base R (v4.2.2) to investigate differences in the median number of reads from small (0.2 and 1.2 µm) vs. large (5.0 and 8.0 µm) filter pore sizes for the different superkingdoms and kingdoms. Here, p-values were adjusted according to the Benjamini–Hochberg approach (Benjamini and Hochberg 1995). The analysis pipeline was implemented using the workflow tool gwf (v1.7.2) (https://github.com/gwforg/gwf) for parallelization, and downstream analyses were computed using R in Jupyter Notebooks and Bash scripts.

Data analysis for metabarcoding

The pipeline MetaBarFlow (v0.1.0) was used to process the metabarcoding data (Sigsgaard et al. 2022). Raw read demultiplexing and primer trimming were done using cutadapt (v3.5) (Martin 2011). Following this, quality trimming and filtering were performed with Sickle (v1.33) (Joshi and Fass 2011) and DADA2 (v1.22.0) (Callahan et al. 2016), respectively. On the Illumina NovaSeq platform, Q scores for base calls have been binned into four categories, and the standard error modeling performed by DADA2 is therefore suboptimal. Thus, we adjusted the parameters (weights and span) of the error model and enforced monotonicity, which provided a much better fit to the data points. Reads that included Ns, contained more than 2 expected errors, and/or were below 40 bp in length were discarded, and reads were truncated at the first instance of a Q score at or below 2. Afterwards, forward and reverse reads were merged, discarding pairs with mismatches or with less than 12 bp of overlap. Chimeric reads, together with PCR and sequencing errors, were removed using DADA2. After quality control and data cleaning, we used BLASTn (v2.12.0+) (Altschul et al. 1990; Benson 2004) to search the resulting amplicon sequence variants (ASVs) against the same local GenBank nt database used for the shotgun sequencing analysis. The search was performed allowing for 1,000 hits per ASV and requiring a minimum of 80% sequence similarity. Hits with less than 100% query coverage were discarded. The R package taxizedb (v0.3.0) (Chamberlain and Arendsee 2021) was used to taxonomically classify each hit. Following Sigsgaard et al. (2022), each ASV was then assigned to the LCA of all matching taxa that overlapped in their range of sequence similarities with that found for the taxon or taxa with the highest sequence similarity.

Further, ASVs with hits below 90% sequence similarity were discarded, and ASVs present in the corresponding PCR blanks were deemed likely contaminants and discarded. We then performed rarefaction curves on the individual PCR reactions and species accumulation curves for each sample to check for sufficiency of sequencing depth and PCR replication. To address the influence of differential sample sequencing depth, we first rarefied the reads in each PCR replicate to the median read number across the four PCR replicates per sample using the R package ROBITools (v0.1). As read depth still differed among samples, we then aggregated the replicates of each sample and rarefied to the minimum read number per sample. Non-metric multidimensional scaling (nMDS) of the results was computed and visualized, as done for shotgun sequencing.

Detections of local and exotic genera

For both shotgun and metabarcoding data, a comparison of Danish genus-level detections within four metazoan groups (fishes (Actinopterygii), mussels (Bivalvia), bristle worms (Polychaeta), and malacostracan crustaceans (Malacostraca)) was carried out using the Danish Species Portal (arter.dk) and the Global Biodiversity Information Facility (gbif.org). Subsequently, for the shotgun sequencing data, pre-LCA BLASTn hits were categorized into Danish and exotic genera. This dataset was used to quantify the occurrence of these genera across samples and replicates (Zenodo; Shotgun II: shotgun.

danish_vs_exotic.read_counts.xlsx) and to compute the overall correlation between the two categories, once accounting for reads containing hits to both Danish and exotic taxa, using Pearson's correlation coefficient (r). Moreover, the BLASTn metrics (percentage of identity, alignment length, number of mismatches, and e-value) were explored to assess the likelihood of exotic identifications. We used Wilcoxon tests across samples and groups to test whether reads associated with Danish taxa reported better metrics than reads associated with exotic taxa (alternative: greater). Resulting p-values were corrected for multiple testing using the Benjamini–Hochberg false discovery rate procedure (Benjamini and Hochberg 1995). To complement significance testing and mitigate the influence of large sample sizes on p-values, we estimated effect sizes using Cliff's delta, calculated with the R package effsize (v0.8.1). Effect sizes were evaluated for each comparison and used to exclude statistically significant results with negligible practical relevance. Statistical significance was defined as an adjusted p-value < 0.05 combined with a Cliff's delta > 0.33, corresponding to a medium to large effect size (Romano et al. 2006). Comparisons involving fewer than 100 hits in either group were excluded from further consideration.

Assessing potential amplification and database bias

To estimate the impact of amplification and database incompleteness bias, we performed an in silico PCR analysis (primers: 18S_allshorts forward and reverse) using Mothur (v1.48.3) (Schloss et al. 2009). For this purpose, we downloaded nucleotide sequences from NCBI using the query "*<taxid>[Organism:exp] AND 18S[All Fields] NOT 28S[All fields] AND 500:3000[SLEN]*" where "<taxid>" corresponded to the taxid of the phyla identified by shotgun and/or metabarcoding. We computed two different Mothur runs using a maximum of six and three mismatches in both forward and reverse primers. The former run was used to estimate the number of 18S sequences covering the barcode region, while the latter was used to estimate the number of sequences that not only covered the barcode region, but were also likely to be amplified by the primers.

Results

Shotgun sequencing output

The shotgun dataset comprised 394.4 million (M) raw reads, of which 96.7% were merged, resulting in 190.6 million paired reads. The sequencing depth ranged from 19.9 to 39.7 million merged reads (mean ± sd: 26,298,125 ± 6,280,913) across the 15 libraries. Notably, the quality assessment revealed consistently high performance, with the proportion of reads with a Q score greater than 30 ranging from 89.9% to 97.6% across libraries.

Querying the NCBI GenBank nt database using BLASTn resulted in hits for 0.7 million to 2.9 million reads per sample (mean ± sd: 1,751,756 ± 673,678.3), showing a similar trend to the initial DNA concentration of the samples, which declined with filter pore size (Fig. 1A). However, the assigned reads accounted for only 0.7% to 4.5% of the total reads from each sample. After applying stringent filtering criteria involving the e-value, percentage of sequence similarity,

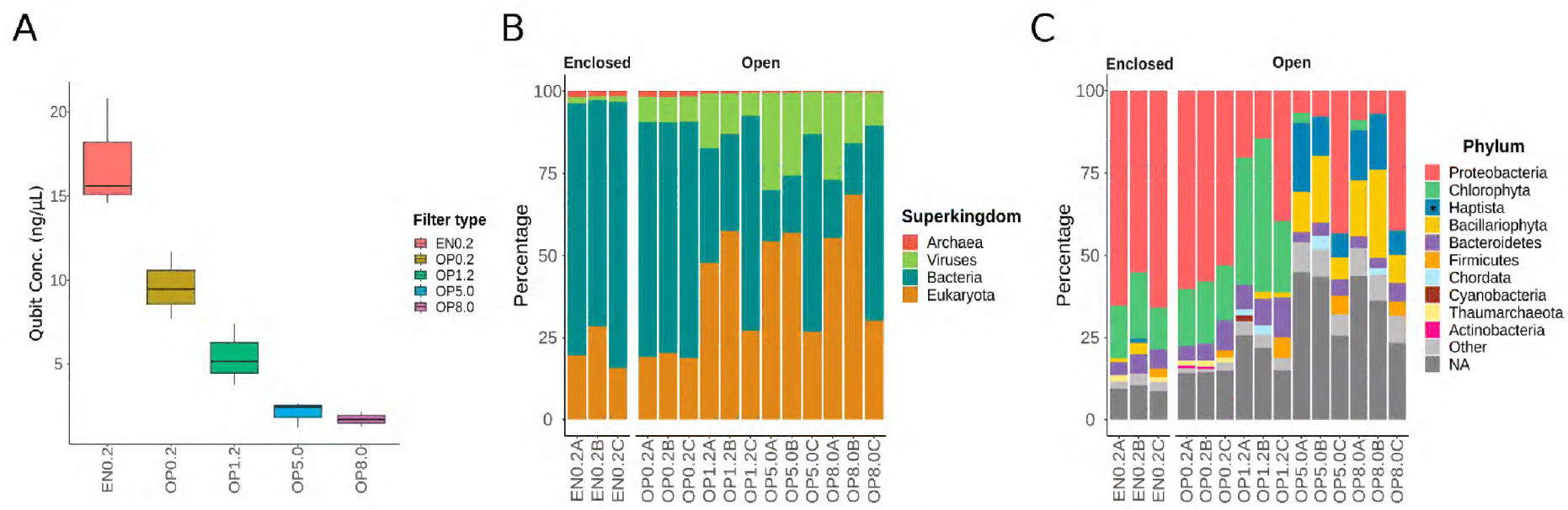

Figure 1. Overview of shotgun data. **A.** Boxplot of initial DNA concentrations (ng/µL) of the eDNA samples measured with Qubit, showing variation among biological replicates; **B.** Stacked barplot showing relative read abundances per sample of the four superkingdoms after rarefaction; **C.** Stacked barplot showing relative read abundances of the five most abundant phyla per sample after rarefaction. Across panels, sample replicates are ordered along the x-axis by increasing pore size, starting with the enclosed filter type (EN) and continuing with the open pore filter type (OP). Pore sizes are in µm. Taxa marked with an asterisk are no longer categorized as a phylum.

query coverage, and query length of database hits, the dataset for each sample was reduced to between 0.2 M and 1.2 M reads (median: 0.6 M reads; mean ± sd: 606,469.3 ± 316,365.9 reads) (Suppl. material 1: table S1). Even for these final reads, species identifications were regularly found to match non-local species, so we decided to collapse our results at the superkingdom and phylum levels (Fig. 1B, C).

Metabarcoding sequencing output

A total of 225.3 M raw reads were obtained from 18S rDNA metabarcoding. Sequencing depth was similar across libraries (mean ± sd: 3,755,348 ± 1,789,722). The initial bioinformatic processing resulted in a total of 119.6 M paired reads (mean ± sd: 1,993,257 ± 1,112,577) (Suppl. material 1: table S2). After merging of paired reads, taxonomic assignment, cleaning, and filtering steps, 55.8 M reads remained. Rarefaction analyses indicated sufficient sequencing depth for each PCR replicate (Suppl. material 1: fig. S2), but also that additional PCR replication would have been necessary to capture the full diversity represented in the eDNA samples (Suppl. material 1: fig. S3). However, the accumulation curves were similar across the three eDNA sample replicates of each sample and across filter pore sizes.

Taxonomic diversity in shotgun sequencing samples

We successfully identified eDNA from 93 phyla, representing all four superkingdoms recognized by NCBI (Archaea, Bacteria, Eukaryota, and Viruses). At the superkingdom level, there were notable differences in community composition between pore sizes (Fig. 1B, C). For bacteria, odds ratios were closer to 1 when comparing the two small pore sizes (0.2 and 1.2 µm) but became much greater than 1 when comparing small to large pore sizes (5.0 and 8.0 µm). This indicates that bacterial reads were strongly enriched in small filters relative to large filters. For Eukaryota, odds ratios were ca. 0.3 when comparing small to

large pore sizes but closer to one when comparing large pore sizes. This indicates that eukaryotic reads were more likely to be detected in large pore sizes, whereas small pore sizes underrepresented eukaryotic taxa relative to other superkingdoms. This pattern is indicative of a progressive decline in the relative read contribution of Bacteria and an increase in Eukaryota with increasing pore size (Suppl. material 1: tables S3, S4; fig. S4). Smaller pore sizes (0.2 and 1.2 µm) yielded higher mean proportions of bacterial compared to eukaryotic reads (63 vs. 28%), while larger pore sizes (5.0 and 8.0 µm) yielded higher mean proportions of eukaryotic compared to bacterial reads (49 vs. 31%). When comparing all superkingdoms across small vs. large filter pore sizes, we found that small pore sizes yielded more reads from Archaea and Bacteria ($p < 0.01$), while large pore sizes yielded more reads from Eukaryota ($p < 0.05$) and Viruses ($p < 0.01$) (Wilcoxon, adjusted p-values) (Fig. 1B; Suppl. material 1: table S5). Similarly, nMDS ordination based on presence/absence of phyla indicated a clear difference between samples, with both axes separating them by filter pore size (Fig. 2, Suppl. material 1: fig. S5).

It is noteworthy that the open filters with a pore size of 0.2 µm yielded similar results to the enclosed filters of the same pore size and that the first two replicates of each filter type (labeled “A” and “B,” respectively) exhibited greater resemblance to each other than either A or B did to the third replicate of each sample (labeled “C”) (Fig. 2). We inspected the taxonomic composition of the individual replicates and identified a large proportion of bacteria in the “C” replicates (Fig. 1C). Here, odds ratios for bacteria were greater than 1 when comparing sample replicate C to A and/or B. Thus, the odds of bacterial reads were greatest in the “C” replicates, while the opposite trend was observed for Eukaryota (Suppl. material 1: table S6). This pattern was most obvious for, but not exclusive to, the largest applied pore sizes. Specifically, we identified a consistently higher abundance of five bacterial phyla (Firmicutes, Fusobacteria, Lentisphaerae, Verrucomicrobia, and Candidatus Melainabacteria) and a lower abundance of two eukaryotic phyla (Discosea and Echinodermata) in the “C” replicates (Fig. 3A).

When analyzing the reads assigned to eukaryotes, only 16.4% of the reads were identified to the kingdom level (Fungi, Metazoa, and Viridiplantae). This percentage corresponded to 5.7% and 0.8% of the reads that were BLAST searched and the paired reads, respectively. Based on this subset of reads, we found that small pore size filters (0.2 and 1.2 µm) yielded more reads from Viridiplantae ($p < 0.001$), while large pore size filters (5.0 and 8.0 µm) yielded more reads from Metazoa ($p < 0.005$), Fungi ($p < 0.005$), and other eukaryotes ($p < 0.001$) (Wilcoxon, adjusted p-values) (Fig. 4A; Suppl. material 1: table S5). Fungi had a very low number of reads compared to the other groups. The relative enrichment or depletion and the abundance of eukaryotic phyla varied depending on filter type and pore size, as shown by the heatmap and boxplot in Fig. 3B, respectively. For algal phyla, Chlorophyta, Bacillariophyta, and Rhodophyta were the most abundant eukaryotic phyla across samples (with Haptista as a remnant phylum). For metazoans, Chordata, Mollusca, Arthropoda, and Echinodermata were consistently the most abundant phyla. The predominant bacterial phyla were Proteobacteria and Bacteroidetes across filter types and pore sizes (Figs 1C, 3A). Other abundant bacterial phyla were Actinobacteria,

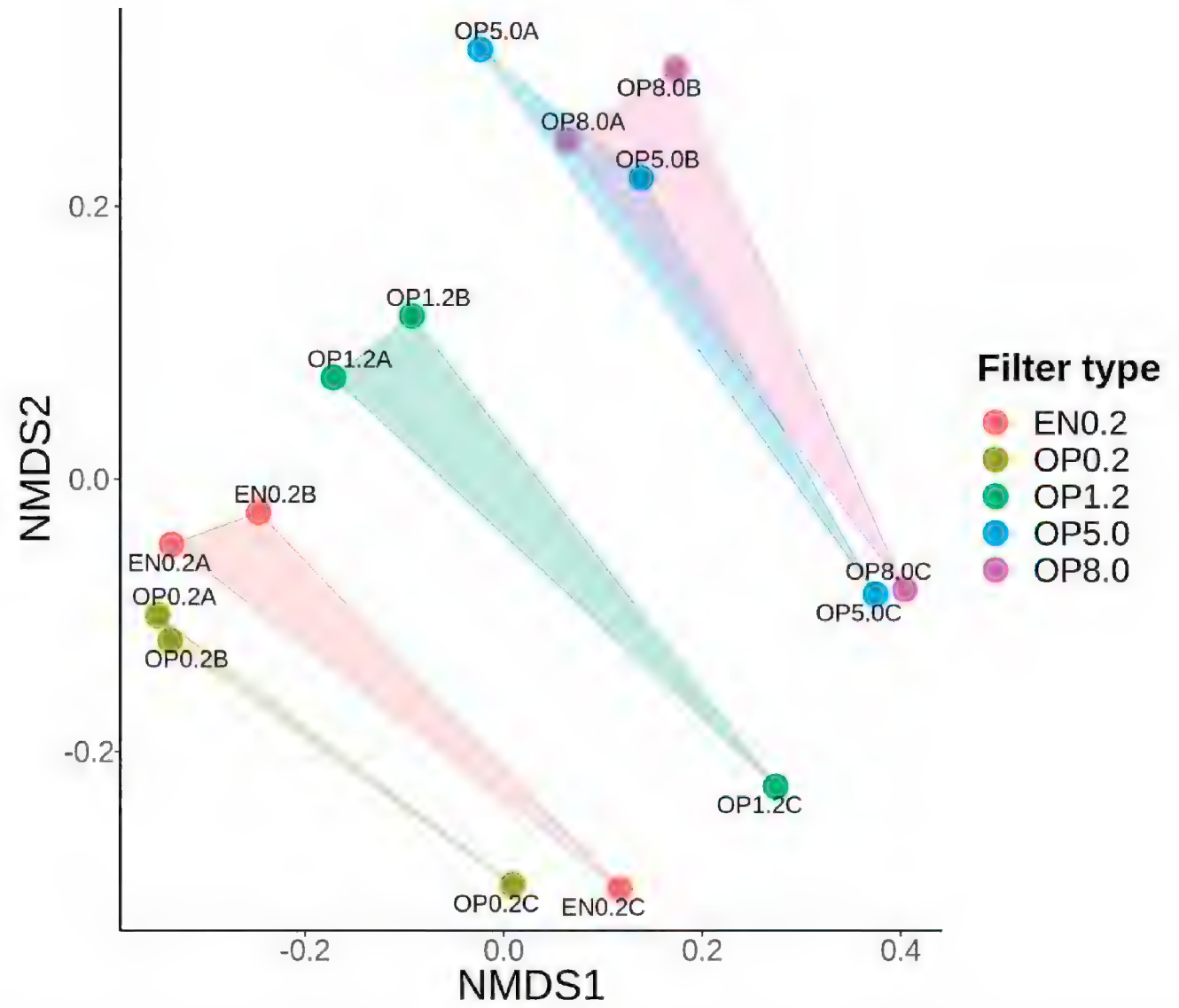

Figure 2. Non-metric multidimensional scaling (nMDS) plot based on Jaccard's dissimilarity index after rarefaction of shotgun sequencing read counts. The x- and y-axes represent the first and second nMDS axes, respectively. Dots represent individual sample replicates colored according to filter type and pore size (µm). Connecting lines between sample replicates of the same pore size represent dispersion within samples.

Cyanobacteria, and Firmicutes. Finally, the only recovered archaeal phyla were Thaumarchaeota, Euryarchaeota, and Crenarchaeota, the latter with an extremely low read count of 18 reads.

Taxonomic diversity in metabarcoding samples

The metabarcoding results were uniquely composed of eukaryotic taxa (Fig. 3C), as expected given that the primers targeted the 18S rRNA region. While most reads could not be assigned to the kingdom level (mean ± sd: 78.4% ± 12.2%), we identified 45 eukaryotic phyla (Fig. 4B, C). After rarefaction of PCR replicates and samples, the majority of the reads that could be assigned to the kingdom level belonged to Metazoa (mean ± sd: 15.6% ± 13.9%), followed by Viridiplantae (mean ± sd: 4.9% ± 4.2%) and then Fungi (mean ± sd: 1.2% ± 0.4%) (Fig. 4B). The nMDS plot based on metabarcoding exhibited a similar trend to the nMDS plot based on shotgun sequencing (Fig. 2), with samples aligning in accordance with increasing pore size along the first axis (Suppl. material 1: figs S6, S7). However, the metabarcoding data did not show a consistent pattern across phyla as observed for shotgun sequencing (Fig. 3B, C). Bacillariophyta, Arthropoda, Ciliophora, Chlorophyta, and Cercozoa, in descending order of abundance, represented the five most abundant phyla across samples when considering both relative read abundance per sample and prevalence, defined as the number of samples in which a phylum was present.

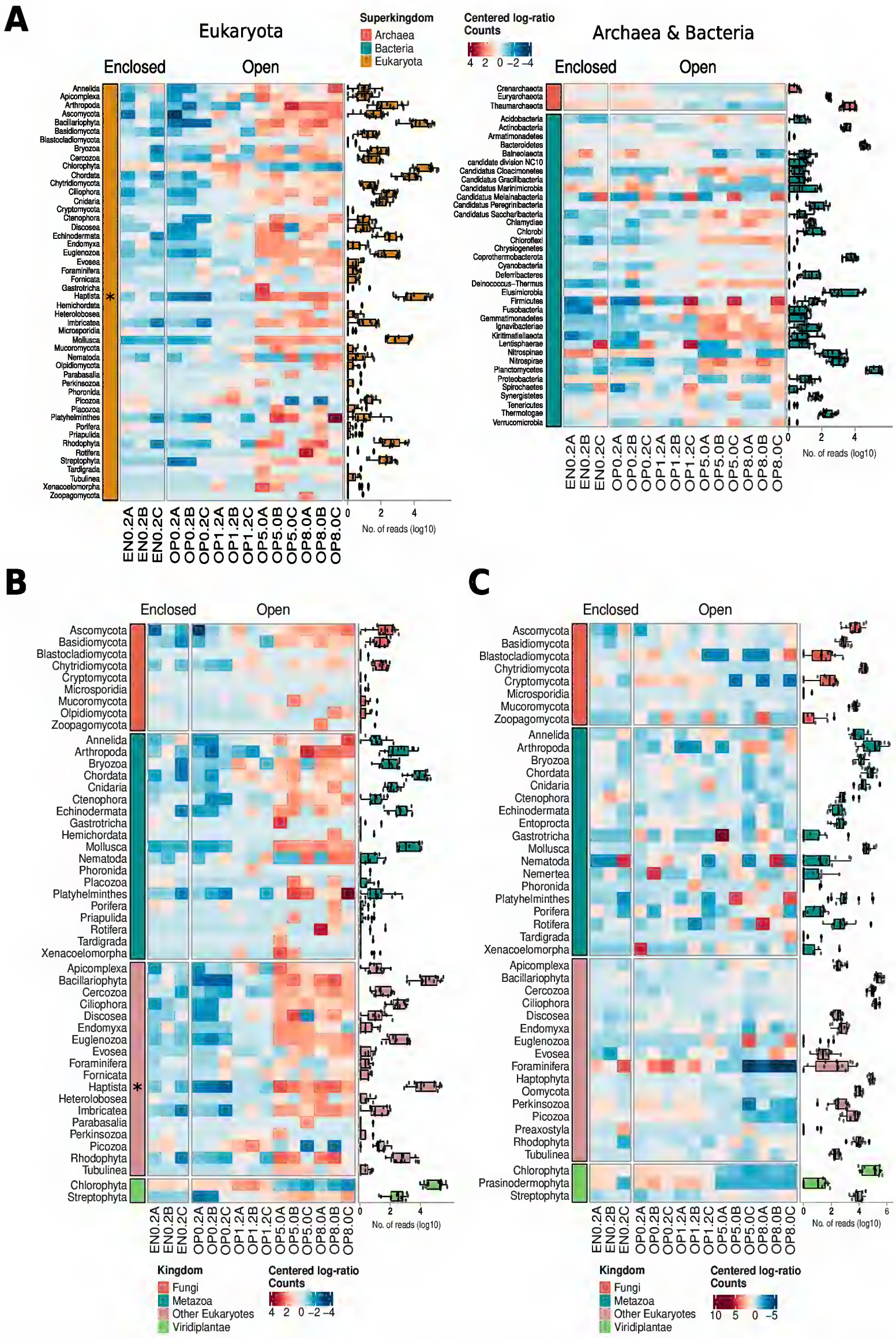

Figure 3. Overview of normalized read count distributions. **A.** Heatmap of zero-handled centered log-ratio (CLR) counts for each phylum detected with shotgun sequencing. The heatmap is split into Eukaryota (left) and Bacteria & Archaea (right); **B.** Heatmap of zero-handled CLR counts for each eukaryotic phylum detected with shotgun sequencing; **C.** Heatmap of zero-handled CLR counts for each eukaryotic phylum detected with metabarcoding. Across panels, a boxplot of read numbers (log10), representing relative abundance per taxon, is shown on the right side (y-axis). Samples are ordered along the x-axis by increasing pore size (µm), starting with the enclosed filter type (EN) and continuing with the open filter type (OP). Phyla on the y-axis are taxonomically ordered by superkingdom or kingdom and listed alphabetically. Values above zero are shown in red, whereas values below zero are shown in blue. Taxa marked with an asterisk are no longer categorized as a phylum.

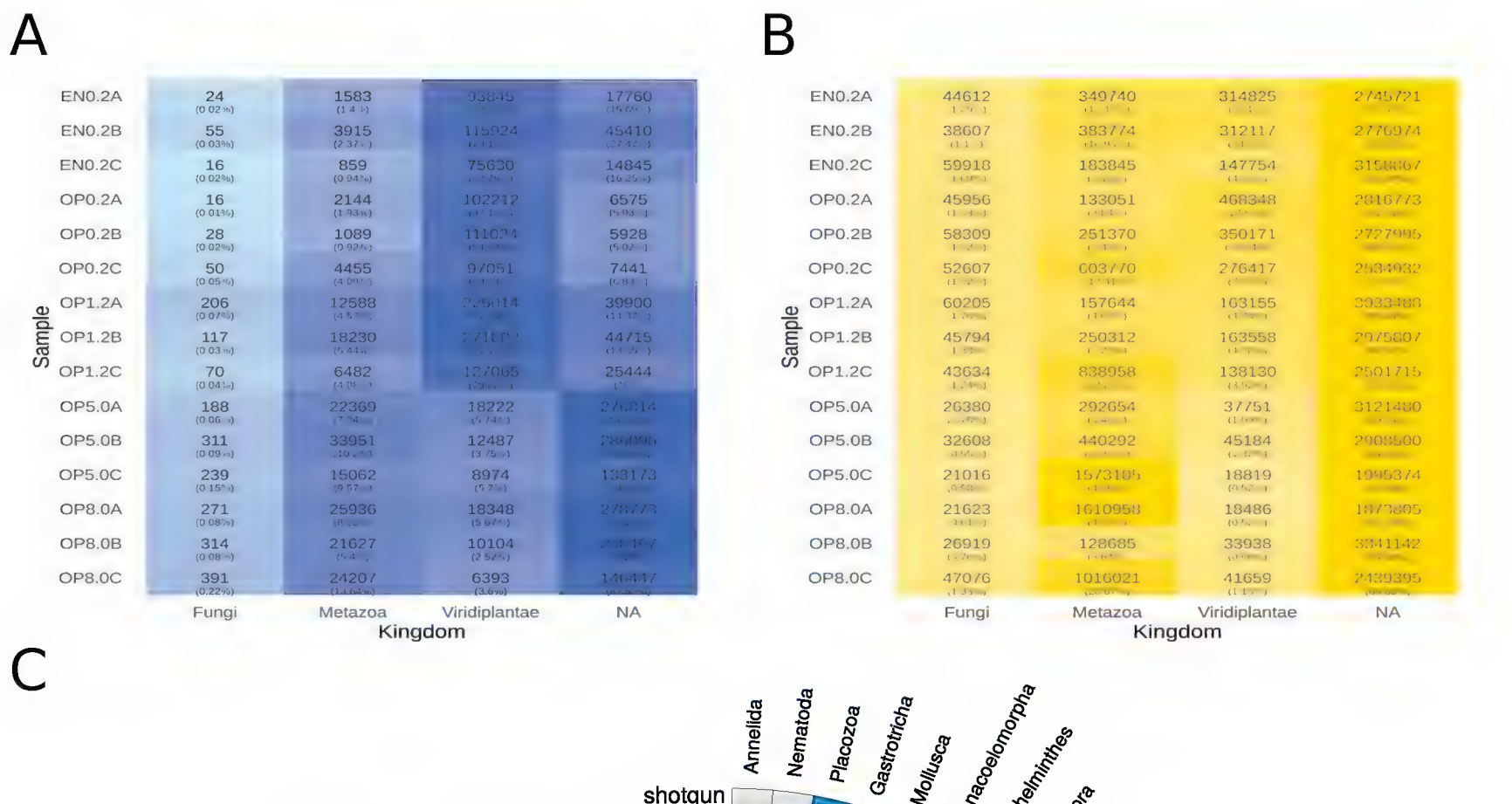

Figure 4. Taxonomic comparison at the kingdom and phylum levels of eukaryotes detected with shotgun sequencing and metabarcoding. **A.** Number and relative abundance (percentages) of taxonomically classified reads after rarefaction per sample and kingdom for shotgun sequencing; **B.** Number and relative abundance of taxonomically classified reads after rarefaction per sample and kingdom for metabarcoding. For both heatmaps, sample replicates are ordered along the y-axis by increasing pore size, starting with the enclosed filter type (EN) and continuing with the open pore filter type (OP). Pore sizes are in µm; **C.** Cladogram of the full set of eukaryotic phyla identified by shotgun sequencing and metabarcoding. Branch colors represent kingdoms (blue, Metazoa; red, Fungi; green, Viridiplantae; pink, other eukaryotes). Three surrounding rings indicate phyla shared between shotgun sequencing and metabarcoding (gray), phyla uniquely detected with metabarcoding (yellow), and phyla uniquely detected with shotgun sequencing (blue). Information in the rings is summarized in a Venn diagram in the upper left corner. Taxa marked with an asterisk are no longer categorized as a phylum.

Comparison of shotgun and metabarcoding data

The analysis at the phylum level revealed a substantial overlap in eukaryotic diversity between the shotgun sequencing and metabarcoding datasets (Fig. 4C). A total of 39 phyla were shared between both methods, while metabarcoding and shotgun sequencing each revealed six and nine unique phyla, respectively. Phyla identified exclusively through metabarcoding encompassed the metazoans Entoprocta and Nemertea and the non-metazoan phyla Haptophyta, Oomycota, Prasinodermophyta, and Preaxostyla. Conversely, phyla found only in the shotgun sequencing data included Fornicata, a group of anaerobic unicellular eukaryotes; the metazoan phyla Hemichordata, Priapulida, and Placozoa; and the non-metazoan phyla Heterolobosea, Imbricatea, Olpidiomycota, and Parabasalia, with Haptista as a remnant phylum. Further, database coverage and amplification of the 45 identified phyla were explored to shed light on potential biases driving the presence of unique phyla in each dataset. Regarding overall nucleotide sequence availability, most phyla showed extensive database coverage (47 of 54 contained >10,000 reads; mean ± sd: 8,872,350 ± 31,326,141), with a few phyla represented by fewer sequences, such as Picozoa, Gastrotricha, and Entoprocta (data available in Zenodo; Shotgun II: Phyla_comparison.xlsx). As expected, the pool of available 18S sequences represented a reduced portion of the total nucleotide sequences available in NCBI (mean ± sd: 8,799 ± 20,345). Despite differences in coverage, the in silico PCR results showed that, on average, 61.3% (sd: ± 22.7%) and 50.2% (sd: ± 26.8%) of the available 18S sequences across phyla could be amplified using the selected primers with six and three mismatches, respectively.

Assessment of Danish vs. exotic taxa at the genus level

To further address the biomonitoring potential of shotgun sequencing, we performed a genus-level comparison of the two methods within four well-known marine metazoan groups for which robust taxonomic knowledge and potentially good data coverage are available, namely fishes (Actinopterygii), mussels (Bivalvia), bristle worms (Polychaeta), and malacostracan crustaceans (Malacostraca) (Suppl. material 1: table S7). For these groups, we counted the number of detected genera known to occur in Denmark and the number of detected genera not known from Denmark, defined as exotic taxa. For the shotgun data, we found that a high to medium proportion of detected genera were Danish for the latter three groups (Polychaeta: 12/13; Malacostraca: 6/7; Bivalvia: 11/18), while for fish genera, fewer than half were Danish (Actinopterygii: 23/59). For the metabarcoding data, the proportions of detected genera known to occur in Denmark were higher than those found for the shotgun data (Polychaeta: 22/23; Malacostraca: 3/4; Bivalvia: 6/6; Actinopterygii: 5/6) (Suppl. material 1: table S7).

We next examined the genus-level composition of Danish and exotic taxa to further evaluate these identifications. First, we compared the number of reads between categories and observed that Danish taxa consistently comprised a higher number of reads, whereas exotic taxa, as well as ambiguous identifications, were always represented by fewer reads (Suppl. material 1: fig. S8). Read abundances for Danish and exotic taxa were weakly to strongly positively correlated for

Actinopterygii, Bivalvia, and Polychaeta (Pearson's r = 0.22, 0.79, and 0.77, respectively) but showed a weak negative correlation for Malacostraca (r = −0.15).

We then tested for differences between Danish and exotic hits across BLAST metrics. No significant differences were detected between categories for alignment length or number of mismatches. In contrast, significant differences were observed for percentage identity and alignment length for fishes and mussels in samples with small pore sizes (EN0.2A, EN0.2B, OP0.2A, OP0.2C, and OP0.2B) and large pore sizes (OP5.0B) (Zenodo; Shotgun II: wilcoxon_cliffs_delta.zip). The differences in mean percentage identity between Danish and exotic hits were small, ranging from 1 to 4%, as were the differences in alignment length, which ranged from 14 to 40 bp.

Discussion

Environmental DNA (eDNA) metabarcoding has gained growing recognition as a valuable tool for biomonitoring of aquatic macro-organisms. Despite the extensive attention dedicated to aquatic eDNA, a gap exists in our understanding of the authentic taxonomic composition captured in seawater samples. This knowledge gap arises from various inherent biases, including, e.g., those pertaining to eDNA sampling strategy, PCR amplification, and/or sequencing methodologies. An alternative, PCR-free (or PCR-limited) sequencing method, shotgun sequencing, has hitherto been deemed inefficient for contemporary eukaryote eDNA studies due to the dominance of bacterial DNA and the lack of reference sequences in public databases, especially outside common barcode regions. However, this technique may offer insights into the composition of eDNA samples that are not attainable with metabarcoding by broadening the taxonomic scope and more accurately reflecting relative abundances of taxa.

In this study, we applied shotgun sequencing to seawater eDNA samples and identified a positive effect of increasing filter pore size on the proportion of eukaryote reads obtained. Overall, compared to small pore sizes, large pore sizes yielded a greater proportion of eukaryotic reads while reducing the proportion of archaeal and bacterial reads. Filters with pore sizes of 5.0 µm and 8.0 µm also captured the largest proportion of eukaryotic phyla. Moreover, we found that shotgun sequencing captured a similar number of eukaryote phyla compared to 18S rDNA metabarcoding. The most obvious caveats of the study are the low proportion of reads that could be taxonomically assigned, the low number of replicates per sample, and the lack of field controls (e.g., sterile water), which are discussed in the following.

Database incompleteness and search strategy

A major drawback of shotgun sequencing data is the large number of reads that cannot be sufficiently identified taxonomically due to database incompleteness. Studies focused on biodiversity assessment using shotgun sequencing of eDNA samples have tackled such issues in various ways. For example, Kjær et al. (2022) included most available references from the NCBI RefSeq repository in their taxonomic assignment, whereas Cowart et al. (2017) relied mainly on databases expected to contain reference data for species from the studied region. Even when leveraging nearly all available plastid and nuclear reference material, shotgun sequencing currently comes up short in comparison with

metabarcoding when assigning taxonomic origins to sequence data. For example, Stat et al. (2017) attained taxonomic assignments for only 14.1% of a total of 22.3 million quality-filtered shotgun reads. In our study, we obtained 190.6 million paired reads in total, but only 26.2 million paired reads (6.6%) of these produced a database match passing our initial alignment requirements. The number of reads was further reduced to 9.0 million reads (2.3%) after additional filtering to increase identification confidence (Suppl. material 1: table S1). The requirements for assigning taxonomy are often based on similarity to matching reference sequences or associated e-values. However, there is currently no consensus on how to properly assign taxonomy when hits exist with low sequence similarities or high e-values. Testing and validation of similarity cutoffs for some barcode regions indicate that across target taxa, similarities of, e.g., 99%, 97%, and 95% are generally sufficient for species-, genus-, and family-level identifications, respectively (e.g., West et al. 2020). With shotgun sequencing, however, such universal criteria cannot be identified, given the large variation in substitution rates across different regions of a genome. Several eDNA and eRNA studies, such as Broman et al. (2021), Jensen et al. (2022), and Pawlowski et al. (2021), have highlighted the potential benefits of utilizing more comprehensive reference data. For example, searching against multiple public sequence databases (e.g., SILVA, BOLD, mmp2, NCBI GenBank and RefSeq, GTDB) offers a chance to increase the taxonomic resolution of eDNA or eRNA results. However, these databases often use different taxonomies, thus requiring extensive bioinformatic work to unite them in a single workflow. Moreover, comparable identification rates between shotgun sequencing and metabarcoding are expected to require the incorporation of substantially expanded reference databases, with an associated increase in computational demands as database size grows. Therefore, it is crucial to weigh the computational effort needed for cross-database searches against potential enhancements in data quality. Furthermore, in addition to the choice of reference database, the identification strategy applied to large shotgun sequencing datasets strongly influences both computational requirements and read retention. In our study, the use of BLAST was computationally intensive and resulted in the exclusion of a substantial fraction of reads. Future analyses might benefit from incorporating alignment-free approaches, such as k-mer–based methods like Kraken2 (Wood et al. 2019), which enable faster and more scalable classification of shotgun sequencing data and might improve the proportion of retained reads.

Comparison between methodologies

Direct comparison between metabarcoding and shotgun sequencing results is very difficult due to general differences between the two methods, more specifically the selectivity and bias of PCR amplification, the different bioinformatic steps necessary for processing the two data types, and the difference in reference database coverage of barcoding regions compared to other genomic loci. Thus, shotgun sequencing is expected to yield a more accurate overall representation of eDNA composition and abundance in a sample, while metabarcoding is generally expected to yield higher taxonomic resolution, although this difference is slight when using a highly generic marker such as the one used in this study.

The large overlap in identified eukaryotic phyla between the two methodologies observed in this study enhances confidence in the shotgun approach, at least at the phylum level. While the results obtained from our in silico PCR amplification analysis lack sufficient taxonomic resolution to be conclusive, they suggest that all phyla were amplifiable and that method-specific detections may be biologically and/or experimentally derived rather than attributable to technical bias. Four of the six phyla that were unique to the metabarcoding dataset (Entoprocta, Nemertea, Prasinodermophyta, and Preaxostyla) contain a moderate number of available reference sequences in the database, while the remaining two (Oomycota and Haptophyta) have substantial reference coverage. Thus, their absence in the shotgun dataset may be attributable either to methodological differences or to unrecorded sequence diversity within these phyla, that is, database incompleteness. On the other hand, the nine phyla unique to the shotgun dataset (Fornicata, Hemichordata, Heterolobosea, Imbricatea, Olpidiomycota, Parabasalia, Placozoa, and Priapulida, including Haptista) represent a combination of taxa with a reduced number of available 18S sequences (i.e., database incompleteness) and lower amplification success with the selected primers (PCR bias). Overall, the shotgun dataset revealed a slightly larger number of phyla than the metabarcoding dataset, supporting the expectation that an almost PCR-free genomic approach provides broader coverage of biodiversity.

The relative abundances of eukaryotic kingdoms varied both between filter pore sizes and between the two methods (Fig. 4A, B). The shotgun sequencing data showed a discernible separation between small pore sizes (0.2 and 1.2 µm) and large pore sizes (5.0 and 8.0 µm) (Fig. 2), which was not present in the metabarcoding data after rarefaction (Suppl. material 1: figs S6, S7). This difference is likely caused by PCR bias but also by the presence of non-eukaryotic taxa in the shotgun data, which underscores the usefulness of shotgun sequencing in discerning the impact of sampling methods on eDNA capture, particularly in relation to filter pore size. A notable difference between enclosed (Sterivex) and open filters of the same pore size (0.2 µm), apart from their difference in filtration area, was a higher relative read abundance of viruses in data from open filters. The cause of this pattern remains speculative, but it may indicate higher levels of contamination or background input, such as aerial transport of DNA, compared to enclosed filters.

Previous studies targeting single species or groups of eukaryotes using PCR have demonstrated an effect of filter pore size on the detection and quantification of these taxa (e.g., Turner et al. 2014; Wacker et al. 2019; Capo et al. 2020; Kumar et al. 2021). The optimal pore size identified in these studies varied and likely depends on, among other factors, the state of the eDNA from the targeted organisms and prevailing environmental conditions. However, it appears to range between approximately 0.2 and 1.5 µm. When targeting eukaryotes using shotgun sequencing, reducing the proportion of non-target DNA at the sampling stage becomes substantially more important, and we found that pore sizes of 5.0 µm or larger were most efficient for this purpose. Inclusion of additional pore sizes, e.g., 0.45 and 3.0 µm, would have allowed closer comparison with previous studies and could, in the future, provide further insight into optimal pore sizes when samples are intended for both metabarcoding and shotgun sequencing.

Caveats and uncertainties

Detection sensitivities of shotgun sequencing might well be improved by increasing sampled water volumes; without PCR amplification, a higher starting amount of eDNA may be needed to detect less abundant organisms or organisms with a lower rate of eDNA "production". However, marine eDNA metabarcoding studies indicate that, at least for metazoans, most of the taxa present at a sampling site can be detected with ca. 1 L water samples, and thus greater spatiotemporal coverage of sampling might be preferred over larger volumes or deeper sequencing (Takahashi et al. 2023).

Interestingly, while the time between samples was minimized as much as possible, the third sample replicate (C) of each pore size was clearly different from the first two sample replicates, although the effect of filter pore size was greater. Jensen et al. (2022) showed short-term changes in coastal eDNA composition at the scale of hours, which is likely due to a combination of biological changes and water movement through currents and waves. The latter, which leads to slightly different water bodies being sampled at a site from hour to hour, could have large effects, as eDNA has been found to be highly spatially heterogeneous (e.g., Bessey et al. 2020). Thus, we reiterate previous recommendations to pay special attention to the number of sample replicates and the spatiotemporal sampling design when performing aquatic eDNA analyses.

A major caveat of the current study was the lack of field controls, which limits the possibility of detecting cross-contamination between samples, contamination from sampling materials, and input from airborne DNA. However, potential uneven contamination of different filter types during sampling is highly unlikely to have given rise to our results. We had three biological replicates per filter type, and our sampling strategy should have distributed any short-term temporal variation in aquatic eDNA composition (Jensen et al. 2022), as well as contamination or "background input", evenly across filter types. Still, a negative field control would have been necessary to rule out the possibility that contamination or eDNA input from the air, which probably influences all eDNA samples (Klepke et al. 2022; Lynggaard et al. 2024), was responsible for the observed differences. "Background input" likely makes up a very small proportion of eDNA data from water samples, as eDNA from the water itself will occur at much higher concentrations, but we stress that eDNA studies should strive to include negative controls at every step of the analysis, normally field, extraction, and PCR controls. Additionally, two further limitations are the use of a single sampling location and the limited number of replicates, both of which constrain the statistical analyses that can be applied. Indeed, our data suggest temporal changes, as replicate C differs from replicates A and B across all filter types; however, the limited design prevents us from formally testing this statistically.

Conclusions and future implications for biomonitoring

In this study, we attempt to elucidate how the qualitative (which organisms are present) and quantitative (how abundant they are) composition of eukaryotic eDNA in a water sample can best be retained in eDNA sequencing data. While the taxonomic coverage of metabarcoding can be greatly expanded by combining several primer sets, such an approach is relatively resource-intensive, and

PCR biases and selectivity make it difficult to infer the initial relative abundances of eDNA from different taxa. Shotgun sequencing is expected to be better suited for this goal, but the predominance of prokaryotes in shotgun sequencing data has until now made the method uncompetitive for detection of eukaryotes. However, we find that increasing filter pore size can significantly decrease the dominance of bacteria in seawater eDNA samples.

Our results have important implications for eDNA-based biomonitoring, as they suggest a simple approach for increasing the efficiency of shotgun sequencing when targeting aquatic eukaryotes and indicate good coverage of diversity at the phylum level compared to metabarcoding. With ongoing reference sequencing efforts, such as the Earth Biogenome Project (Lewin et al. 2022), on the horizon, a larger fraction of eDNA-derived shotgun data will likely be identifiable in the future. These new references will both increase the proportion of taxonomically assignable reads in the data and significantly increase the taxonomic resolution of reads from currently underrepresented nuclear DNA regions. In addition, substantial improvements in bioinformatic approaches for analyzing shotgun data in ancient eDNA studies (Kjær et al. 2022) will be valuable to consider in future studies of contemporary eDNA.

Currently, taxonomic identification at the family, order, or even class level is very speculative for shotgun sequencing, and best practice for eDNA biomonitoring at lower taxonomic levels (e.g., species or genera) will therefore likely continue to be metabarcoding, using, e.g., COI or 12S for animals and ITS2 for plants and fungi, for some time. Notably, a comparison of our results at the genus level for four well-studied metazoan groups showed similar ratios of native, identification likely correct, to exotic, identification likely incorrect, genera for shotgun sequencing compared to 18S metabarcoding. It is important to stress that we regard hits to exotic genera as incorrect identifications due to database incompleteness for taxa and genetic regions, rather than actual detections of overlooked marine taxa in Denmark. Although the latter is a possibility, particularly for bivalves and crustaceans that could be transported with ballast water as larvae, we consider it much more unlikely than the former issue related to database coverage. The metabarcoding data showed much lower proportions of exotic species, reflecting higher accuracy, for 18S, which is not normally used for metabarcoding at lower taxonomic levels, and using, e.g., COI metabarcoding would most likely increase accuracy further compared to shotgun sequencing. However, for our shotgun sequencing results, it is notable that we observed a substantially lower number of reads for exotic taxa compared to Danish taxa, especially for the larger pore sizes where metazoans are enriched (Suppl. material 1: fig. S8). This suggests that a cutoff value based on read numbers, although difficult to determine and likely to vary between studies, could almost exclusively separate Danish taxa from exotic, incorrect identifications. This would make shotgun sequencing a strong competitor to metabarcoding for biomonitoring of metazoans. We found that for fishes, the native proportions of detected genera were particularly low for shotgun sequencing. Fishes have relatively high coverage of genomes in public reference databases, suggesting that shotgun sequencing requires further optimization before being applicable in biomonitoring. We leave it to future studies to test shotgun approaches for biomonitoring at appropriate taxonomic levels. This will require better curated genomic reference databases and more thoroughly tested bioinformatic

approaches than those used here (e.g., k-mer-based searches). Considering (i) the low taxonomic selectivity and bias, (ii) the genome-wide coverage, and (iii) the continuous and rapid increase in cost-efficiency of sequencing and in database completeness, we believe that shotgun sequencing of water samples should be further developed for more accurate and comprehensive aquatic biodiversity monitoring.

Acknowledgments

We thank Katrine Bay Jensen for performing the DNA extractions. We thank GenomeDK, Center for Genome Analysis and Personalized Medicine, Aarhus University, as well as Servicio de Genómica y Secuenciación Masiva (GENGS) of Centro de Biología Molecular Severo Ochoa (CBM Severo Ochoa), for providing computational resources and support that contributed to these research results.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Use of AI

No use of AI was reported.

Funding

This study was financially supported by the Faculty of Natural Sciences, Aarhus University (grant 27744), the Carlsberg Foundation (grant CF18-0949), and the Velux Foundations (grant 21517).

Author contributions

PFT conceived the idea for the study. PFT and EES carried out eDNA sampling. EES performed the laboratory work together with the technical staff. AGR and EES analyzed the data with input from PFT and MRJ. EES and MRJ provided the pipeline used to perform the metabarcoding computations. AGR led the manuscript writing with assistance and input from all authors. All authors contributed to the manuscript drafts and approved the final version.

Author ORCIDs

Adrián Gómez Repollés https://orcid.org/0009-0002-6335-321X
Eva Egelyng Sigsgaard https://orcid.org/0000-0002-9396-1550
Mads Reinholdt Jensen https://orcid.org/0000-0001-8240-1083
Philip Francis Thomsen https://orcid.org/0000-0002-9867-4366

Data availability

The raw sequencing data are available in ENA/NCBI under the BioProject PRJEB104040. The results from the various analyses, including the intermediate files required to demultiplex the 18S raw sequencing data and to run the scripts, are available in Zenodo (Shotgun

I: https://doi.org/10.5281/zenodo.17256866, Shotgun II: https://doi.org/10.5281/zenodo.17426358; Metabarcoding I: https://doi.org/10.5281/zenodo.17635072). Finally, the conda environments, pipelines, notebooks, and scripts are available on GitHub (Shotgun and downstream analysis: https://github.com/adriangeerre/eDNAPoreSize; Metabarcoding: https://github.com/evaegelyng/MetaBarFlow).

References

Agersnap S, Sigsgaard EE, Jensen MR, Avila MDP, Carl H, Møller PR, Krøs SL, Knudsen SW, Wisz MS, Thomsen PF (2022) A National Scale "BioBlitz" Using Citizen Science and eDNA Metabarcoding for Monitoring Coastal Marine Fish. Frontiers in Marine Science 9. https://doi.org/10.3389/fmars.2022.824100

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215(3): 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andruszkiewicz EA, Starks HA, Chavez FP, Sassoubre LM, Block BA, Boehm AB (2017) Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. PLoS One 12(4): e0176343. https://doi.org/10.1371/journal.pone.0176343

Antich A, Palacín C, Cebrian E, Golo R, Wangensteen OS, Turon X (2021) Marine biomonitoring with eDNA: Can metabarcoding of water samples cut it as a tool for surveying benthic communities? Molecular Ecology 30(13): 3175–3188. https://doi.org/10.1111/mec.15641

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B, Statistical Methodology 57(1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benson DA (2004) GenBank. Nucleic Acids Research 33(Database issue): D34–D38. https://doi.org/10.1093/nar/gki063

Bessey C, Jarman SN, Berry O, Olsen YS, Bunce M, Simpson T, Power M, McLaughlin J, Edgar GJ, Keesing J (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA 2(4): 493–504. https://doi.org/10.1002/edn3.74

Bowers H, Pochon X, von Ammon U, Gemmell N, Stanton JA, Jeunen GJ, Sherman C, Zaiko A (2021) Towards the Optimization of eDNA/eRNA Sampling Technologies for Marine Biosecurity Surveillance. Water (Basel) 13(8): 1113. https://doi.org/10.3390/w13081113

Broman E, Bonaglia S, Norkko A, Creer S, Nascimento FJA (2021) High throughput shotgun sequencing of eRNA reveals taxonomic and derived functional shifts across a benthic productivity gradient. Molecular Ecology 30(13): 3023–3039. https://doi.org/10.1111/mec.15561

Bushnell B (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13(7): 581–583. https://doi.org/10.1038/nmeth.3869

Capo E, Spong G, Königsson H, Byström P (2020) Effects of filtration methods and water volume on the quantification of brown trout (Salmo trutta) and Arctic char (Salvelinus alpinus) eDNA concentrations via droplet digital PCR. Environmental DNA 2(2): 152–160. https://doi.org/10.1002/edn3.52

Chamberlain S, Arendsee Z (2021) taxizedb: Tools for Working with "Taxonomic" Databases (0.3.0). CRAN.

Coissac E (2012) OligoTag: A Program for Designing Sets of Tags for Next-Generation Sequencing of Multiplexed Samples. In Pompanon F, Bonin A (Eds) Data Production

and Analysis in Population Genomics: Methods and Protocols. Humana Press, 13–31. https://doi.org/10.1007/978-1-61779-870-2_2

Cowart DA, Murphy KR, Cheng CCH (2017) Metagenomic sequencing of environmental DNA reveals marine faunal assemblages from the West Antarctic Peninsula. Marine Genomics 37: 148–160. https://doi.org/10.1016/j.margen.2017.11.003

De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. Molecular Ecology Resources 14(2): 306–323. https://doi.org/10.1111/1755-0998.12188

de Cárcer DA, Denman SE, McSweeney C, Morrison M (2011) Evaluation of Subsampling-Based Normalization Strategies for Tagged High-Throughput Sequencing Data Sets from Gut Microbiomes. Applied and Environmental Microbiology 77(24): 8795–8798. https://doi.org/10.1128/AEM.05491-11

Fowler S, Roush R, Wise J (2017) Concepts of Biology. OpenStax, Rice University.

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in Ecology and Evolution 7(11): 1299–1307. https://doi.org/10.1111/2041-210X.12595

Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics (Oxford, England) 32(18): 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Guardiola M, Uriz MJ, Taberlet P, Coissac E, Wangensteen OS, Turon X (2015) Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons. PLoS One 10(10): e0139633. https://doi.org/10.1371/journal.pone.0139633

Jensen MR, Sigsgaard EE, Liu S, Manica A, Bach SS, Hansen MM, Møller PR, Thomsen PF (2021) Genome scale target capture of mitochondrial and nuclear environmental DNA from water samples. Molecular Ecology Resources 21(3): 690–702. https://doi.org/10.1111/1755-0998.13293

Jensen MR, Sigsgaard EE, Ávila M de P, Agersnap S, Brenner-Larsen W, Sengupta ME, Xing Y, Krag MA, Knudsen SW, Carl H, Møller PR, Thomsen PF (2022) Short-term temporal variation of coastal marine eDNA. Environmental DNA 4(4): 747–762. https://doi.org/10.1002/edn3.285

Jo T, Takao K, Minamoto T (2022) Linking the state of environmental DNA to its application for biomonitoring and stock assessment: Targeting mitochondrial/nuclear genes, and different DNA fragment lengths and particle sizes. Environmental DNA 4(2): 271–283. https://doi.org/10.1002/edn3.253

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (1.33).

Kjær KH, Winther Pedersen M, De Sanctis B, De Cahsan B, Korneliussen TS, Michelsen CS, Sand KK, Jelavić S, Ruter AH, Schmidt AMA, Kjeldsen KK, Tesakov AS, Snowball I, Gosse JC, Alsos IG, Wang Y, Dockter C, Rasmussen M, Jørgensen ME, Skadhauge B, Prohaska A, Kristensen JÅ, Bjerager M, Allentoft ME, Coissac E, PhyloNorway Consortium, Rouillard A, Simakova A, Fernandez-Guerra A, Bowler C, Macias-Fauria M, Vinner L, Welch JJ, Hidy AJ, Sikora M, Collins MJ, Durbin R, Larsen NK, Willerslev E (2022) A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. Nature 612(7939): 283–291. https://doi.org/10.1038/s41586-022-05453-y

Klepke MJ, Sigsgaard EE, Jensen MR, Olsen K, Thomsen PF (2022) Accumulation and diversity of airborne, eukaryotic environmental DNA. Environmental DNA 4(6): 1323–1339. https://doi.org/10.1002/edn3.340

Kumar G, Farrell E, Reaume AM, Eble JA, Gaither MR (2021) One size does not fit all: Tuning eDNA protocols for high- and low-turbidity water sampling. Environmental DNA. https://doi.org/10.1002/edn3.235

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, Blaxter ML, Cai J, Caperello ND, Carlson K, Castilla-Rubio JC, Chaw SM, Chen L, Childers AK, Coddington JA, et al. (2022) The Earth BioGenome Project 2020: Starting the clock. Proceedings of the National Academy of Sciences of the United States of America 119(4): e2115635118. https://doi.org/10.1073/pnas.2115635118

Lynggaard C, Frøslev TG, Johnson MS, Olsen MT, Bohmann K (2024) Airborne environmental DNA captures terrestrial vertebrate diversity in nature. Molecular Ecology Resources 24(1). https://doi.org/10.1111/1755-0998.13840

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17(1): 10. https://doi.org/10.14806/ej.17.1.200

Mauvisseau Q, Harper LR, Sander M, Hanner RH, Kleyer H, Deiner K (2022) The Multiple States of Environmental DNA and What Is Known about Their Persistence in Aquatic Environments. Environmental Science & Technology 56(9): 5322–5333. https://doi.org/10.1021/acs.est.1c07638

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, Kondoh M, Iwasaki W (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. Royal Society Open Science 2(7): 150088. https://doi.org/10.1098/rsos.150088

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2020) vegan: Community Ecology Package (2.5-7). CRAN.

Pan W, Byrne-Steele M, Wang C, Lu S, Clemmons S, Zahorchak RJ, Han J (2014) DNA polymerase preference determines PCR priming efficiency. BMC Biotechnology 14(1): 10. https://doi.org/10.1186/1472-6750-14-10

Pascoal F, Tomasino MP, Piredda R, Quero GM, Torgo L, Poulain J, Galand PE, Fuhrman JA, Mitchell A, Tinta T, Turk Dermastia T, Fernandez-Guerra A, Vezzi A, Logares R, Malfatti F, Endo H, Dąbrowska AM, De Pascale F, Sánchez P, Henry N, Fosso B, Wilson B, Toshchakov S, Ferrant GK, Grigorov I, Vieira FRJ, Costa R, Pesant S, Magalhães C (2023) Inter-comparison of marine microbiome sampling protocols. ISME Communications 3(1). https://doi.org/10.1038/s43705-023-00278-w

Pawlowski J, Bonin A, Boyer F, Cordier T, Taberlet P (2021) Environmental DNA for biomonitoring. Molecular Ecology 30(13): 2931–2936. https://doi.org/10.1111/mec.16023

Pinfield R, Dillane E, Runge AKW, Evans A, Mirimin L, Niemann J, Reed TE, Reid DG, Rogan E, Samarra FIP, Sigsgaard EE, Foote AD (2019) False negative detections from environmental DNA collected in the presence of large numbers of killer whales (Orcinus orca). Environmental DNA 1(4): 316–328. https://doi.org/10.1002/edn3.32

Power H, Takahashi M, Jarman S, Berry O (2023) What is environmental DNA? Environmental DNA 5(6): 1743–1758. https://doi.org/10.1002/edn3.497

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nature Biotechnology 35(9): 833–844. https://doi.org/10.1038/nbt.3935

Romano J, Kromrey JD, Coraggio J, Skowronek J (2006) Appropriate Statistics for Ordinal Level Data PDF level of Measurement Effect Size. In annual meeting of the Florida Association of Institutional Research. https://www.scribd.com/document/45715000/Appropriate-Statistics-for-Ordinal-Level-Data

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology 75(23): 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated – reducing sequence to sample misidentifications in metabarcoding studies. Molecular Ecology Resources 15(6): 1289–1303. https://doi.org/10.1111/1755-0998.12402

Shirazi S, Meyer RS, Shapiro B (2021) Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. Ecology and Evolution 11(22): 15766–15779. https://doi.org/10.1002/ece3.8239

Sigsgaard EE, Nielsen IB, Carl H, Krag MA, Knudsen SW, Xing Y, Holm-Hansen TH, Møller PR, Thomsen PF (2017) Seawater environmental DNA reflects seasonality of a coastal fish community. Marine Biology 164(6): 128. https://doi.org/10.1007/s00227-017-3147-4

Sigsgaard EE, Torquato F, Frøslev TG, Moore ABM, Sørensen JM, Range P, Ben-Hamadou R, Bach SS, Møller PR, Thomsen PF (2020) Using vertebrate environmental DNA from seawater in biomonitoring of marine habitats. Conservation Biology: The Journal of the Society for Conservation Biology 34(3): 697–710. https://doi.org/10.1111/cobi.13437

Sigsgaard EE, Soraggi S, Jensen MR, Repollés AG, Thomassen EE, Thomsen PF (2022) MetaBarFlow (0.3.0). https://doi.org/10.5281/zenodo.6006700

Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, David LA (2021) Measuring and mitigating PCR bias in microbiota datasets. PLoS Computational Biology 17(7): e1009113. https://doi.org/10.1371/journal.pcbi.1009113

Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M (2017) Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. Scientific Reports 7(1): 12240. https://doi.org/10.1038/s41598-017-12501-5

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198767220.001.0001

Takahashi M, Saccò M, Kestel JH, Nester G, Campbell MA, Van Der Heyde M, Heydenrych MJ, Juszkiewicz DJ, Nevill P, Dawkins KL, Bessey C, Fernandes K, Miller H, Power M, Mousavi-Derazmahalleh M, Newton JP, White NE, Richards ZT, Allentoft ME (2023) Aquatic environmental DNA: A review of the macro-organismal biomonitoring revolution. The Science of the Total Environment 873: 162322. https://doi.org/10.1016/j.scitotenv.2023.162322

Thomsen PF, Willerslev E (2014) Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation 183: 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Tillotson MD, Kelly RP, Duda JJ, Hoy M, Kralj J, Quinn TP (2018) Concentrations of environmental DNA (eDNA) reflect spawning salmon abundance at fine spatial and temporal scales. Biological Conservation 220: 1–11. https://doi.org/10.1016/j.biocon.2018.01.030

Turner CR, Uy KL, Everhart RC (2014) Fish environmental DNA is more concentrated in aquatic sediments than surface water. Biological Conservation 183: 93–102. https://doi.org/10.1016/j.biocon.2014.11.017

Wacker S, Fossøy F, Larsen BM, Brandsegg H, Sivertsgård R, Karlsson S (2019) Downstream transport and seasonal variation in freshwater pearl mussel (Margaritifera margaritifera) eDNA concentration. Environmental DNA 1(1): 64–73. https://doi.org/10.1002/edn3.10

West KM, Stat M, Harvey ES, Skepper CL, DiBattista JD, Richards ZT, Travers MJ, Newman SJ, Bunce M (2020) eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. Molecular Ecology 29(6): 1069–1086. https://doi.org/10.1111/mec.15382

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, et al. (2019) Welcome to the Tidyverse. Journal of Open Source Software 4(43): 1686. https://doi.org/10.21105/joss.01686

Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. Genome Biology 20(1): 257. https://doi.org/10.1186/s13059-019-1891-0

Supplementary material 1

Supplementary tables and figures

Authors: Adrián Gómez-Repollés, Eva Egelyng Sigsgaard, Mads Reinholdt Jensen, Philip Francis Thomsen

Data type: docx

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.10.164232suppl1